心理科学进展 2021, Vol. 29, No. 2, 225–237
Advances in Psychological Science

https://dx.doi.org/10.3724/SP.J.1042.2021.00225

·研究方法(Research Method)·

# 问卷调查中被试不认真作答的控制与识别*

钟晓钰 李铭尧 李凌艳

(北京师范大学中国基础教育质量监测协同创新中心，北京 100875)

**摘 要** 问卷调查是心理与教育领域十分常见的数据收集方法，而被试的不认真作答可能导致问卷数据失真。回顾已有研究发现：(a)不认真作答可以从外在作答模式和内在产生原因两个方向进行定义；(b)不认真作答的常见事前控制方法主要包括降低任务难度以及提高被试作答动机两大类；(c)事后识别方法主要包括嵌入识别量表、作答模式识别、反应时识别三大类。今后的研究中应基于作答机制的研究优化与开发控制方法，检验作答识别方法的跨情境适用性并开发新方法，并对局部不认真的识别与处理进行更深入的探讨。

**关键词** 不认真作答，数据清理，无效数据，调查问卷设计

**分类号** B841

在科学研究中，当研究者将所要调查的内容具体化为一系列有机联系的可测指标，进而编制成问题表格或簿本(刘蔚华，陈远，1991)，旨在测量人的行为或态度时，就形成了问卷(车文博，2001)。问卷调查是社会科学研究中十分常见的数据收集方式，但是，通过这种方式获得的数据容易包含较多测量误差，因此在基于数据建模、推断、决策之前需要对其进行筛选，以识别和纠正这些不正确的结果(Huang et al., 2012)。

在这些误差中，不认真作答是既常见、又往往因难以处理而被忽视的因素之一。研究表明，在大多数问卷调查中不认真作答的发生率从1% (Gough & Bradley, 1996)到30% (Burns et al., 2014)不等。不认真作答会污染数据结果，大大降低数据的真实性，如不加处理，可能会掩盖有意义的结果、产生虚假结果(Curran, 2016; Maniaci & Rogge, 2014)。其影响主要包括：第一，影响测量工具的信效度(DeSimone et al., 2018; Kam & Meyer, 2015; Zijlstra et al., 2011)，例如，单维量表中的反向表述题越容易从正向表述题中脱离成单独的维度(Woods, 2006)。第二，形成随机(random)数据或奇异值(outlier)，进而影响随后的推断与决策(Barge & Gehlbach, 2012; Huang et al., 2015; Zijlstra et al., 2011)，例如影响百分等级评分(Zijlstra et al., 2011)、夸大或缩小变量间的相关等(Credé, 2010; Holtzman & Donnellan, 2017; Huang et al., 2015; Schneider et al., 2018)。

随着电子问卷使用的愈加广泛(Evans & Mathur, 2005; Lloyd & Devine, 2010)，问卷提交的便利性(Johnson, 2005)、作答的匿名性(Meade & Craig, 2012)、作答环境的不可控(Barge & Gehlbach, 2012; Carrier et al., 2009; Meade & Craig, 2012)、主试与被试互动的减少(Francavilla et al., 2019; Johnson, 2005; Ward & Meade, 2018; Zhang & Conrad, 2018)等原因会大大增加不认真作答的风险(Ward & Pond, 2015)。基于此，本文对相关研究进行系统概括和总结，以期提高研究者与实践者对问卷不认真作答的重视，并为其选用控制与识别方法提供参考：首先梳理了国外研究中不认真作答的相关概念以明确其范畴，之后分别总结不认真作答的控制与识别技术，最后对未来研究的方向做了展望。

## 1 不认真作答的相关概念

“不认真作答”这一概念在英文语境中尚无统

收稿日期：2020-04-27

* 国家社会科学基金一般项目“基于大数据循证的学校治理现代化研究” (20BGL234)。

通信作者：李凌艳, E-mail: lilingyan@bnu.edu.cn

一的术语，且不同研究使用的术语存在微妙的差别，这些术语主要有两个侧重方向：外在作答模式和内在产生原因。

### 1.1 外在作答模式

不认真作答的其中一类概念着重描述外显结果，即作答模式(response pattern)，多指李克特式量表中的选项分布。例如被研究者广泛采用的术语随机作答(random responding)，即被试在问卷中随机地勾选(Beach, 1989; Berry et al., 1992; Marjanovic et al., 2015)。但也有研究者指出，不认真作答可能呈现出非随机的模式(Meade & Craig, 2012)，例如直线作答(straight-lining & nondifferentiation) (Curran, 2016; Fang et al., 2016; Huang et al., 2012; Meade & Craig, 2012)，或按照无意义的规律选择答案等(Dunn et al., 2018)。此外，Grau 等人(2019)也发现不认真作答与特定作答风格(response style)存在一定程度的重合。各作答模式示例如图 1 所示。

这些研究直观地描述了不认真作答外显的作答模式，同时认可这些模式产生的原因是被试的不努力、不认真。这种“不努力”恰恰是不认真作答与社会称许性反应(social desirability responding)的差别——社会称许性反应也可能表现为特定的作答风格(He & van de Vijver, 2013, 2015a, 2015b, 2016)，但它并非减少了答题过程中的认知负荷，反而“需要额外认知努力” (Grau et al., 2019; Maniaci & Rogge, 2014; McGrath et al., 2010; Meade & Craig, 2012)。然而，由于不认真作答模式复杂多样难以穷举，仅从模式表现上的描述会造成对该概念的窄化。

### 1.2 内在产生原因

为了避免上述的窄化，有研究者在定义时更侧重不认真作答的产生原因。Krosnick (1991)认为被试作答的努力程度是一个从理想最大值(optimization)到完全不努力的连续体，任务难度、被试能力和被试作答动机共同影响了被试在这一连续体上的位置。Zhang (2013)将这一理论进一步细化，区分了努力程度的理想最大值(a)、可达到最大值(attainable maximum) (b)、实际值(actual) (c)三个节点。其中任务难度和被试能力决定了可达到最大值(b)的位置，被试作答动机决定了实际值(c)的位置(如图 2 所示)。不认真作答则是被试因为作答动机较低，从而出现不遵循问卷的指导语、没有精准地理解题目内容、没有提供准确回答的行为(Bowling et al., 2016; Huang et al., 2012; Meade & Craig, 2012)。这类概念包括缺乏努力的作答(insufficient effort responding) (Huang et al., 2012)、粗心的作答(careless responding) (Grau et al., 2019; Johnson, 2005; Meade & Craig, 2012)、非卷入的作答(disengaged responding) (Soland et al., 2019)、逃避行为(shirking behavior) (Fang et al., 2016)、不专心(inattention) (Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012)、令自我满意的作答行为(satisficing behaviors) (Anduiza &

| | | | | | |
|---|---|---|---|---|---|
| Q1 | **1** | 2 | 3 | 4 | 5 |
| Q2 | 1 | 2 | **3** | 4 | 5 |
| Q3 | 1 | 2 | **3** | 4 | 5 |
| Q4 | 1 | **2** | 3 | 4 | 5 |
| Q5 | 1 | 2 | 3 | **4** | 5 |
| Q6 | 1 | 2 | 3 | 4 | **5** |
| Q7 | 1 | 2 | 3 | **4** | 5 |
| Q8 | 1 | 2 | 3 | 4 | **5** |
| Q9 | 1 | **2** | 3 | 4 | 5 |
| Q10 | **1** | 2 | 3 | 4 | 5 |

随机作答

| | | | | | |
|---|---|---|---|---|---|
| Q1 | **1** | 2 | 3 | 4 | 5 |
| Q2 | 1 | **2** | 3 | 4 | 5 |
| Q3 | 1 | 2 | **3** | 4 | 5 |
| Q4 | 1 | 2 | 3 | **4** | 5 |
| Q5 | 1 | 2 | 3 | 4 | **5** |
| Q6 | 1 | 2 | 3 | 4 | **5** |
| Q7 | 1 | 2 | 3 | **4** | 5 |
| Q8 | 1 | 2 | **3** | 4 | 5 |
| Q9 | 1 | **2** | 3 | 4 | 5 |
| Q10 | **1** | 2 | 3 | 4 | 5 |

按照无意义的规律选择答案

| | | | | | |
|---|---|---|---|---|---|
| Q1 | 1 | 2 | 3 | 4 | **5** |
| Q2 | 1 | 2 | 3 | 4 | **5** |
| Q3 | 1 | 2 | 3 | 4 | **5** |
| Q4 | 1 | 2 | 3 | 4 | **5** |
| Q5 | 1 | 2 | 3 | 4 | **5** |
| Q6 | 1 | 2 | 3 | 4 | **5** |
| Q7 | 1 | 2 | 3 | 4 | **5** |
| Q8 | 1 | 2 | 3 | 4 | **5** |
| Q9 | 1 | 2 | 3 | 4 | **5** |
| Q10 | 1 | 2 | 3 | 4 | **5** |

直线作答

| | | | | | |
|---|---|---|---|---|---|
| Q1 | 1 | 2 | 3 | 4 | **5** |
| Q2 | 1 | 2 | 3 | 4 | **5** |
| Q3 | 1 | 2 | 3 | 4 | **5** |
| Q4 | 1 | 2 | 3 | **4** | 5 |
| Q5 | 1 | 2 | 3 | 4 | **5** |
| Q6 | 1 | 2 | 3 | 4 | **5** |
| Q7 | 1 | 2 | 3 | 4 | **5** |
| Q8 | 1 | 2 | 3 | 4 | **5** |
| Q9 | 1 | 2 | 3 | 4 | **5** |
| Q10 | 1 | 2 | 3 | 4 | **5** |

特定作答风格
(如, 默许肯定风格)

图 1 各类作答模式示例

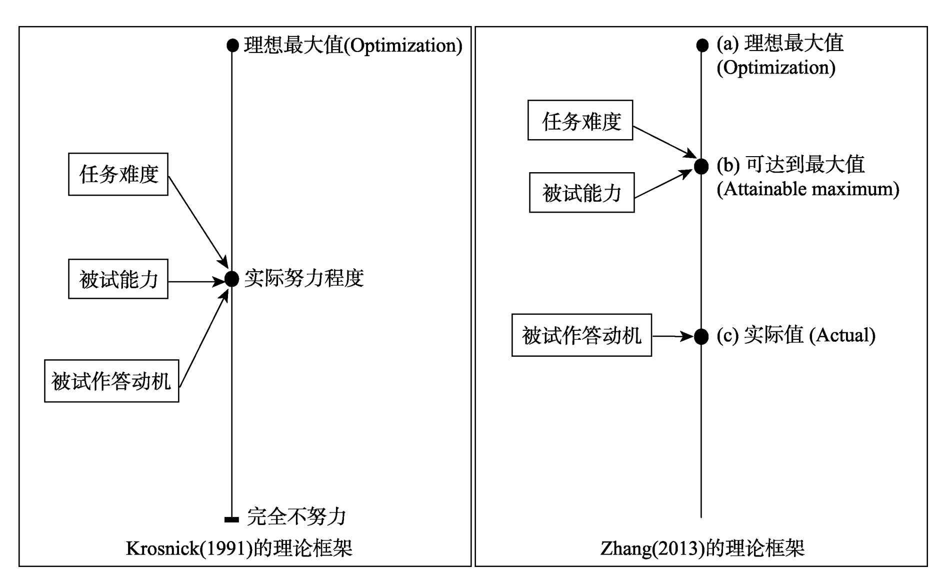



图 2　Krosnick (1991)及 Zhang (2013)的理论框架
图片来源: Zhang, 2013

Galais, 2017; Barge & Gehlbach, 2012; Zhang & Conrad, 2018)等等。

以上两类术语从两个侧面对不认真作答进行了描述或定义，二者并不冲突，且在相互补充之下，丰富了不认真作答的内涵。基于这些术语，有研究者提出，不认真作答可以被定义为，个体在作答问卷过程中因动机不足而表现出的不遵从题目要求，或未仔细阅读题目内容便做出回答的作答模式，其外显形式包括随机作答、直线作答等等(Huang et al., 2012)。

## 2　不认真作答的事前控制

事前控制指在编制问卷或施测时通过某种方法阻止或者减少被试作答不认真的现象。控制方法主要分为两大类，第一，降低任务难度以提高努力程度的可达到最大值，常见手段为调整问卷表述与长度。第二，提高作答动机从而提高努力程度的实际值，常见手段有施加外部奖惩、要求被试承诺认真作答以及提供反馈增加社会互动。

### 2.1　降低任务难度

依据 Zhang (2013)的理论，任务难度会影响被试的努力程度中可达到的最大值。而在调查问卷中，降低任务难度一方面体现在为被试提供清晰、合适、易于理解的指导语和题目表述，进而减轻被试认知加工负担(García, 2011; Rousseau & Ennis, 2013)；另一方面体现在缩短问卷，降低被试的疲劳感。若问卷过长，被试在作答至问卷中间或靠后的位置时，可能精力不足，注意无法持续集中(卫旭华, 张亮花, 2019)，或产生厌烦感和枯燥感，出现不认真作答的现象(Baer et al., 1997; Berry et al., 1992)。实证研究证明，单次填答较长的调查问卷会对数据质量产生负面影响(Nguyen, 2017)。因此有研究者建议，当被测量构念是定义清晰的单维构念且非研究中的核心构念时，尤其是在大样本、时间受限的研究中，对同一构念的测量可以采用单题项的方式缩短问卷，以提高数据收集的有效性(卫旭华, 张亮花, 2019)。

### 2.2　提高作答动机

不认真作答的事前控制更多着力于激发被试的作答动机，以提高被试努力程度的实际值。当被试不愿意或不认为自己应当对结果负责时，就不会持谨慎的态度，实际认真程度会远低于最大可能的认真程度(Ward & Meade, 2018)。而提高被试作答动机主要包含以下几种方式：

1)施加外部奖惩。由于大多数问卷调查对被试而言是低利害或无趣的(卫旭华, 张亮花, 2019)，所以问卷自身无法使被试保持较高的作答动机，因而需要一些外部的奖励或警告。其中，外部奖励(如被试费)是吸引被试填答问卷的常见手段。但当奖励的目的性过强时，被试可能会为了获得奖

励而随意应付调查(Barge & Gehlbach, 2012; Maniaci & Rogge, 2014)。因此除了奖励，警告也是有必要的。警告通常出现在指导语中，例如告知被试调查结束后研究者会采用统计手段评估作答质量，将有问题的数据剔除，或将数据质量反馈给被试，甚至对不认真作答的被试有所惩罚(如不支付被试费等)。有研究表明，警告对控制不认真作答有显著的效果(Huang et al., 2012; Ward & Pond, 2015)。

2)要求被试承诺认真作答。一旦人们明确承诺一个行动或立场，他们倾向于以与承诺相一致的方式行事(Cialdini, 2001)。但直接的承诺未必能达到理想的效果，Cibelli (2017)在实验中要求被试承诺“认真思考、努力回忆、花时间填答”，以增加被试的责任感。实验结果显示承诺在提高作答质量上作用有限，仅能使得被试在难题(如，主观题)上付出更多努力。此外，被试填答时往往会无视指导语，因此有研究者提出可以通过在问卷前设置指示题(instructional manipulation checks)提醒被试认真作答，被试只有正确回答指示题才可以继续填答问卷。Oppenheimer 等人(2009)发现，这种方式使得不认真作答的情况整体得到改善。

3)提供反馈增加社会互动。这类方法主要针对电子问卷。首先，在被试作答过快或者连续选择同一选项时出现弹窗提示，能够提升数据质量(Cibelli, 2017; Zhang, 2013; Zhang & Conrad, 2018)。其次，在电子问卷中，缺少与主试之间的社会互动被认为是被试难以维持填答动机和认知努力的原因(Fang et al., 2014; Meade & Craig, 2012)，因此提高社会互动也是降低不认真作答的思路之一。Ward 和 Pond (2015)通过在电子问卷中放置“虚拟人(virtual humans)”的方式模拟纸笔测验时被试与主试之间的社会互动，提升被试的注意力和责任感。实验证明当警告的指导语与监督的“虚拟人”同时存在时，不认真作答在被试中的发生率显著降低。但 Francavilla 等人(2019)的研究结果显示，“虚拟人”的作用有限，实验组的被试仅在少数指标上表现更好。再次，有研究进一步分析了反馈中“社会性”的作用，即在反馈弹窗中用人脸图片替代黄色感叹号图标，但结果显示这两种方法之间没有显著差异(Zhang, 2013; Zhang & Conrad, 2018)。此外，弹窗信息和“虚拟人”也存在分散被试注意力的潜在风险(Ward & Pond, 2015)。

## 3 不认真作答的事后识别

事前控制能减少不认真作答的发生，但不能完全避免，因此有必要在数据收集之后，对原始数据中仍存在的不认真作答数据进行事后识别与剔除。已有研究开发出许多事后识别的方法，按证据来源可划分为三类：嵌入识别量表、作答模式识别及反应时识别。

### 3.1 嵌入识别量表

嵌入识别量表也被称作主动筛查法(proactive approaches) (Dunn et al., 2018)或者直接筛查法(direct screening methods) (Desimone et al., 2015)，其基本原理是在原问卷中嵌入识别量表，反映被试不认真作答程度。识别量表题主要有三类：陷阱题(bogus items)、指示题和自我汇报题(self-report)。

1)陷阱题，即正确答案显而易见的题目。例如“我于 2 月 30 日出生” (Huang et al., 2012)、“我已经周游了世界 92 次” (Dunn et al., 2018)等。这类题虽然与周围题目一样采用李克特 5 点或 7 点计分的方式询问被试的同意程度，但只有“非常不同意”是合理的。如果被试多次在这类题目上选择其它选项，则会被认为不认真。

2)指示题，即要求被试按照题干的指示进行操作的题目。例如“请在本题选择第二个选项” (Anduiza & Galais, 2017)、“请跳过本题” (Maniaci & Rogge, 2014)、“请点击屏幕下方的小圆圈” (Oppenheimer et al., 2009)。如果被试多次出现不按题干指示作答的情况，则会被认为不认真。

3)自我报告题，即直接询问被试对自己认真努力程度的主观判断。例如“我并没有太在意这些问题的实际含义”、“我回答问题的时候很粗心” (Huang et al., 2012)。这种识别方法简单而直接，如果被试承认自己作答不认真，则研究者也会将其标记。

识别量表简单、直观，是最为普遍的识别方法，但其也存在两方面的问题。一方面，不认真作答者未必完全不看题目，若这类量表题和问卷主体内容毫无关联，被试仅需动用极少认知资源就能注意到，因此该方法只能最低程度地识别不认真作答。其次，嵌入量表题目过多可能会激怒认真作答的被试(Costa Jr & McCrae, 2008; Curran,

2016; Meade & Craig, 2012)。

### 3.2 作答模式识别

依据作答模式识别, 也称反应性筛查(reactive approaches)。此类方法在数据收集之后对被试的作答模式进行分析, 计算识别指标, 表示被试不认真作答的程度(Meade & Craig, 2012)。识别逻辑主要有个体一致性(individual consistency)分析和奇异值分析两种。

#### 3.2.1 个体一致性分析

在李克特量表中, 不认真作答的常见表现形式为随机作答和直线作答(Curran, 2016; Maniaci & Rogge, 2014; Meade & Craig, 2012; Revilla & Ochoa, 2015)。因此, 这类指标假定, 如果被试在各题目上的选项分布过于随机, 或过于一致, 则表明其没有认真作答(Barge & Gehlbach, 2012; Marjanovic et al., 2015)。常见指标包括长线系数(long string index)、作答标准差(inter-item standard deviation, ISD)、个人信度(individual reliability)、正/反向题目对相关。

1)长线系数, 即连续选择某一选项的最长个数, 该指标对直线作答十分敏感(Meade & Craig, 2012)。例如, 当被试在一个 10 题的 4 点计分量表中作答模式为[1, 1, 1, 2, 1, 2, 2, 3, 4, 4], 则连续选择同一选项的个数分别为[3, 1, 1, 2, 1, 2], 其中最大值 3 即为长线系数, 均值 1.67 亦可作为衡量不认真作答的指标; 也有研究者采用每个选项对应的长线系数(Costa Jr & McCrae, 2008; Huang et al., 2012), 在本例中, 答案 1~4 对应的长线系数分别为[3, 2, 1, 2]。

2)作答标准差, 又称个人作答变异系数(intra-individual response variability index) (Curran, 2016; Dunn et al., 2018; Marjanovic et al., 2015)。其计算公式是:

$$ISD_i = \sqrt{\frac{\sum_{g=1}^{k}(X_{ig} - \bar{X}_i)^2}{(k-1)}}\ .$$

其中 $ISD_i$ 表示被试 i 的作答标准差, $X_{ig}$ 是被试 i 在第 g 题上的得分, $\bar{X}_i$ 是被试 i 所有题目的均分, k 是题目总数。当被试作答过于随机时, 其单个维度中的 ISD 会异常大; 而被试作答过于一致时, 其整个问卷的 ISD 会异常小(Dunn et al., 2018; Marjanovic et al., 2015)。研究者建议整个问卷题量在 25~150, 单个维度内题目大于 5 时更适合计算 ISD (Barge & Gehlbach, 2012; Dunn et al., 2018)。

3)个人信度。利用个人信度测量不认真作答有以下前提假设: 每一个子量表都只测量一个心理构念; 不认真作答的被试采取的方式是随机作答(Curran, 2016)。个人信度最常见的指标是奇偶一致系数(even-odd consistency) (Huang et al., 2012; Jackson, 1976, 1977; Johnson, 2005; Meade & Craig, 2012)。其计算过程是先将整个问卷分为若干个子量表, 再分别计算每个子量表的奇数项和偶数项的平均值, 求奇数项平均值组成的向量和偶数项平均值组成的向量之间的相关, 最后用斯皮尔曼−布朗公式进行校正。Jackson (1977)建议当奇偶一致系数小于 0.30 的时候, 可以认为该被试很大概率作答不认真。Curran (2016)提出一种新计算方法, 称作重复取样个人信度(Resampled Individual Reliability, RIR)系数, 与奇偶一致系数逻辑相同, 但通过重复不断的抽样(resampling and bootstrapping)获得尽可能多的分半样本以得到更稳健的结果。

4)正/反向题目对相关, 是指量表中意义相同或者意义相反的两个题目组成的题目对之间的相关。其中构建题目对的方法有两种: 一种称为“语义上的(semantic)题目对”, 是在题目设计之初制定的; 另一种称为“心理测量上的(psychometric)题目对”, 是通过数据驱动的方式进行构建的(Curran, 2016), 依据 Johnson (2005)的建议, 可以利用已采集的数据计算题目间的两两相关, 相关系数在 0.60 以上的题目对可以构建成心理测量上的正/反向题目对。而个人作答的认真程度可以通过正/反向题目对得分的相关值体现。

尽管个体一致性的各识别指标在理解与计算上相对直观, 但被试作答的一致性程度受问卷内容、长度和形式等因素影响, 这使得各识别指标很难制定跨问卷的临界值(cutoff), 且在有些情况下这些指标的识别效果有限。例如, 利用长线系数识别不认真作答明显在短问卷中有较大局限性(Curran, 2016); 再者, 在某些内容领域(如态度、适应性)的调查中, 得分分布并非正态, 而常常呈现偏态(牟智佳, 2017; 王俪嘉, 朱德全, 2009; 姚成 等, 2012; 郑云翔 等, 2018), 这也就意味着被试选择很多“非常同意”也是正常的。又如, 当问卷中存在反向表述的题目时, 对分数大小敏感的个

人信度、作答标准差等指标的使用也需要更加谨慎(Curran, 2016)。

### 3.2.2 奇异值分析

奇异值分析的基本假设是“任何给定样本中的大多数被试都在认真思考并答题” (Curran, 2016)。因此当个人作答模式偏离群体程度过大时，可以认为该被试作答不认真。常见的指标有：马氏距离(Mahalanobis distance)、被试拟合系数(individual respondent's goodness-of-fit score, $R_{GF}$)、人总相关系数(person-total correlation)、个人拟合指数(person-fit statistics)中的Guttman错误个数(Guttman error)、U3指数、$l_Z$指数、神经网络(neural network)算法中的自动编码器(autoencoder)等等。

1)马氏距离(Mahalanobis, 1936)，这是一个常用的多变量奇异值识别指标，且在大多统计软件中可以直接计算。定义

$$\mathrm{MD}_i=\sqrt{(x_i-\mu)^{\mathrm{T}}S^{-1}(x_i-\mu)}$$

为第 $i$ 个样本的马氏距离。其中 $x_i=(x_{i1},\cdots,x_{ik})^T$ 为样本 $i$ 在 $k$ 个维度上的得分；$\mu=(\mu_1,\cdots,\mu_k)^T$ 是 $x$ 的期望；$S$ 是 $x$ 的协方差矩阵。Meade 和 Craig (2012)通过模拟结果发现，马氏距离是一个强大的探测不认真作答的指标。Velleman 和 Welsch (1981)建议用也可以用杠杆值 $h_{ii}=\frac{1}{n-1}MD^2+\frac{1}{n}$ 判断奇异值，以 $\frac{2k}{n}$ 或 $\frac{3k}{n}$ 临界值，其中 $k$ 为变量个数，$n$ 为样本量。

2)被试拟合系数(Kountur, 2016)，其计算公式如下：

$$R_{GFi}=\sum_{g=1}^{k}\frac{(X_{gi}-\bar{X}_g)^2}{\bar{X}_g}$$

其中 $R_{GFi}$ 是代表作答认真程度的被试拟合系数，$X_{gi}$ 是被试 $i$ 在第 $g$ 道题目上的得分。$\bar{X}_g$ 是所有被试在第 $g$ 道题目上得分的均值。被试拟合系数反映了某个作答与整体作答之间的偏差，当被试偏离整体的程度越大时，被试拟合系数的数值越大。

3)人总相关系数(Curran, 2016)，即某被试作答模式 $X$ 与其他所有人作答模式 $M$ 的相关系数 $\rho_{XM}$，其中 $M=E(X)$。如果人总相关系数较低，则说明该被试的作答模式与总体有较大的背离，可能是该被试作答不认真。

4)个人拟合指数，在成就测验领域使用个人拟合指数来识别异常个体已经得到广泛认可，其逻辑是比较分数的观测分布和理想分布的拟合程度(Meijer & Sijtsma, 2001)。这一逻辑近年来也被迁移至问卷调查不认真反应的识别中。其中，理想分布需要使用群体作答模式数据进行构建，因此通过个人拟合指数进行不认真作答识别也需要假定大部分人是认真作答者(Meijer & Sijtsma, 2001; Wang & Xu, 2015)。常见用于识别不认真作答的个人拟合指数有多级计分中 Guttman 错误(Guttman error)的个数 $G^P$ (Emons, 2008; Guttman, 1944, 1950)及 $G^P$ 的标准化形式 $G_N^P$ (Emons, 2008)、U3指数(van der Flier, 1980)的多级计分版本 $U3^P$ (Emons, 2008)、$l_Z$ 指数的多级计分版本 $l_Z^P$ (Melipillán, 2019)等等。

①Guttman错误个数。Guttman模型(Guttman model)的基本逻辑是被试应该更容易在简单题目上得分。它最开始被用于成就测验(二级计分)，例如，将测验中所有题目按正确率 $\pi_g$ 从大到小降序排列，如果被试在靠前的相对简单题上没有得分，而靠后的相对难题上得分了，则不符合 Guttman 模型，犯了 Guttman 错误。Guttman 错误越多，数据越异常。定义 Guttman 错误的个数 G 为：

$$\mathrm{G}=\sum_{h,e}X_{nh}(1-X_{ne})$$

$X_{nh}$ 表示被试在两道题中相对难的题目上的得分(1为正确，0为错误)，$X_{ne}$ 表示被试在两道题中相对简单的题目上的得分。

实际上，Guttman模型也可以扩展到多级计分中，进而可以在李克特量表式问卷中计算 $G^P$ (Emons, 2008; Niessen et al., 2016)。即基于优势模型(dominance model)的测量理论，被试的特质水平越高，越容易打高分，也就是越容易跳过前一个选项(如，非常不同意)而选择后一个选项(如，比较不同意)。此时，可以用计算“测验正确率”的逻辑计算每一个题目的每一个选项的通过概率 $\pi_g$。Emon (2008)同时提出 $G^P$ 的标准化版本 $G_N^P$，便于跨情境对比。

②U3指数。U3指数是一种常见且具有较好检验力(power)的非参数个人拟合指数(Karabatsos, 2003)。它同样源于成就测验，在成就测验中非参数个人拟合指数的一般表达式为：

$$G_i = \frac{\sum_{g=1}^{r} w_g - \sum_{g=1}^{k} X_g w_g}{\sum_{g=1}^{r} w_g - \sum_{g=k-r+1}^{k} w_g}$$

其中 $i$ 为被试编号, $g$ 为题目序号, $k$ 为题目总数($g$ = 1, ⋯, $k$), $X_g$ 为被试在第 $g$ 题上的得分, $r$ 为被试答对的题目数(Meijer & Sijtsma, 2001)。$w_g$ 为适应性函数, 在不同的个人拟合指数中 $w_g$ 的计算有所不同, 而在 U3 指数中 $w_g = \ln\left(\frac{\pi_g}{1-\pi_g}\right)$。$G_i$ 的绝对值越小, 异常程度越低, 当 $G_i$ 为 0 时, 数据符合 Guttman 模型。与 Guttman 错误个数一样, 当用题目的选项通过率 $\pi_g$ 代替正确率时, U3 指数同样可用于多级计分的量表中(Emons, 2008)。

③ $l_Z$ 指数。Levine 和 Rubin (1979)提出似然估计指标(log-likelihood fit)是个人拟合中研究应用最为广泛的指数。$l$ 指数属于参数个人拟合指数, 表示个人得分模式和 IRT 模型拟合的理想模式之间的差异, $l_Z$ 指数即 $l$ 的标准化形式(Drasgow et al., 1985)。$l$ 指数计算公式为:

$$l = \sum_{g=1}^{k} \{X_g lnP_g(\theta) + (1-X_g)\ln[1-P_g(\theta)]\}$$

在二级计分(如成就测验)中 $P_g(\theta)$ 表示能力 θ 的被试在题目 g 上答对的概率; 在多级计分中则记为 $P_{x_g}(\theta)$, 表示通过题目 $g$ 的选项 $x_g$ 的概率(Melipillán, 2019)。$l$ 指数与 $l_Z$ 指数越小, 异常程度越大。

5)自动编码器。自动编码器是非监督神经网络中常用于识别高维度奇异值的方法, 被广泛运用于工程学领域, Melipillán (2019)将其用于识别问卷的不认真作答。自动编码器的原理是将数据先降维编码, 再升维解码, 比较生成数据与原始数据的差距。对于奇异值而言, 其生成数据和原始数据的差距一般较大。在事先设置合适阈值的情况下, 即可标记奇异值。Melipillán 的研究中, 利用自动编码器的方法经过 4 次迭代识别奇异值的整体效果优于利用 $l_Z$ 指数识别。

然而, 任何奇异值指标的效果都非常依赖整个样本的性质, 即奇异值分析只能说明该被试的作答是否偏离群体, 无法断定偏离群体的原因, 这使得采用奇异值分析识别问卷中的不认真作答值得商榷。首先, 低利害调查中不认真作答的比例可能非常大(不同于奇异值分析常用的考试领域, 异常作答情况较少), 被奇异值指标标记的异常被试很可能是认真作答者, 而不是数量可观的不认真作答者。其次, 个体在各个题目上得分不同本属正常现象, 当用这种差异判定个体是否认真作答时, 可能会把一部分认真作答的极端个体排除。最后, 作假等其他因素也可能造成数据异常, 因此通过奇异值指标标记的异常被试不一定是不认真造成的。此外, 这些奇异值指标也有各自的优势和缺陷, 例如马氏距离虽然可以在大多数统计软件上直接计算, 但其要求数据服从多元正态分布, 而问卷中的数据常常难以满足这一前提(Niessen et al., 2016); 又如, 个人拟合指数在题量较少时虽有较高敏感度, 但其计算基于优势模型的理论假设, 可能不符合态度调查的认知过程; 再如, 神经网络算法的结果难以解释, 且较难保证跨情境的稳定性。

### 3.3 反应时识别

一般认为, 当作答时间非常短、被试在回答问题之前完成基本阅读都是不可能的情况下, 其给出的回答难以代表其真实想法(Huang et al., 2012)。反应时阈值的设定有四种方法: 依据经验设定、观察反应时分布图像、结合其他数据质量指标设定以及进行实验预试。

依据经验设定的反应时阈值可以分为绝对标准和相对标准, 其中绝对标准中运用最为广泛的阈值是 Huang 等人(2012)“有根据地猜测”的题均 2 秒(Curran, 2016; Soland et al., 2019)。也有研究设定相对标准, Höhne 和 Schlosser (2018)总结了过往研究中五个相对标准(如表 1 所示)。

第二种常见的方法是通过观测反应时分布图像来确定阈值。例如, 假设认真答题的被试需要至少 5 秒钟的时间来阅读、理解和回答题目, 那么正常作答的情况下, 时间分布应该大于 5 秒; 但是不认真作答的被试可能不需要 5 秒就能完成回答。在这种情况下, 整个群体的反应时应该呈双峰分布(如图 3)。最初几秒内出现的是不认真作答的“尖峰”, 之后是正常作答行为的反应时(Wise, 2017; Wise & Demars, 2006; Wise & Kong, 2005)。

第三种方法是利用其他识别指标(如前述长线系数等)与反应时进行关联以帮助确定阈值, 或者验证已有阈值合理性。Soland 等人(2019)在世界经济合作与发展组织(Organization for Economic

表 1 离群反应时上下阈值(Höhne & Schlosser, 2018)

| 文献来源 | 阈值下限 | 阈值上限 |
|---|---|---|
| Mayerl (2013) | Mean − (2 × *SD*) | Mean + (2 × *SD*) |
| Schnell (1994) | $Q_{.50} - (1.5 \times IQR)$ | $Q_{.50} + (1.5 \times IQR)$ |
| Hoaglin et al. (2000) | $Q_{.50} - (1.5 \times (Q_{.50} - Q_{.25}))$ | $Q_{.50} + (1.5 \times (Q_{.75} - Q_{.50}))$ |
| Hoaglin et al. (2000) | $Q_{.50} - (3 \times (Q_{.50} - Q_{.25}))$ | $Q_{.50} + (3 \times (Q_{.75} - Q_{.50}))$ |
| Lenzner et al. (2010) | $Q_{.01}$ | $Q_{.99}$ |

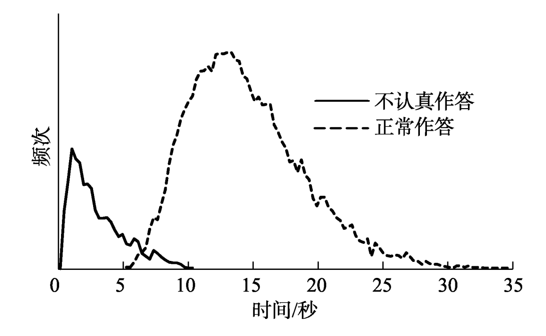



图 3 快速猜测(不认真)作答和正常作答的反应时理论分布

Co-operation and Development, OECD)的学校测试数据中利用该种策略，首先按照一定经验准则将题均反应时分成若干区间，并分别计算每个区间内被试的长线系数、反向题目对相关、EFA 第二特征根的大小、自我效能问卷得分与相应学科成就测验得分的相关等若干指标。结果发现当题均反应时小于 2 秒时，以上指标的表现都较差。

最后一种方法为事先进行实验预试，Huang 等人(2012)在研究中首先通过实验室的指导语将被试控制为认真作答组与不认真作答组，并获得两组被试包括反应时在内的各项指标数据；接着他们再固定特异性(specificity)为 95%和 99%，得到各指标的阈值与对应的敏感度(sensitivity)；最后将从实验中获得的各指标阈值运用于问卷调查的筛查中。

反应时由于不受被试作答模式影响，还可以细化到题目水平进行评估，所以较多研究都发现反应时是有效的不认真作答识别指标(Huang et al., 2012; Wise & Kong, 2005)。但反应时也存在一定缺陷：第一，反应时数据获取困难，只有电子问卷才可能记录。第二，和其他识别指标一样，反应时能否有效区分正常作答被试和不认真作答被试，取决于不认真作答被试在该指标上和正常作答被试的重叠程度，当不认真作答的分布偏离正常分布不大时，识别效果将会降低(Curran, 2016)；而这一点在问卷调查中格外明显，因为不同于认知测验，问卷题目即使认真阅读与思考也无需花费太长时间，这使得通过反应时进行数据清理可能存在较多“误杀”，所以也有研究认为认知测验中快速猜测行为与正常答题行为的反应时理论分布(双峰分布)难以在调查问卷中推广使用(Soland et al., 2019)。第三，反应时的增加并不一定意味着数据质量的增加(Yan & Tourangeau, 2008), Meade 和 Craig (2012)认为反应时和数据质量之间存在非线性关系，作答非常快的被试是不认真的，但作答非常慢的被试，一旦超过既定的阈值，也可能被认为是不认真的。比如在网络调查中，反应时过长可能是因为被试在和他人聊天、看电视或听音乐(Barge & Gehlbach, 2012; Börger, 2016)。

## 4 讨论与展望

不认真作答是调查问卷中常见的噪音源，本文首先梳理了不认真作答的相关概念，接着综述了不认真作答的各种事前控制与事后识别方法。下面探讨问卷不认真作答领域中，有待研究者探索和解决的问题。

### 4.1 基于不认真作答的产生机制，优化与开发控制方法

已有研究发现，调整问卷表述或长度、奖励、警告、弹窗提醒、“虚拟人”、承诺及前置指示题均在一定程度上有助于减少不认真作答的发生，但这些方法也可能产生副作用甚至反作用，如外部激励可能导致被试态度更为散漫，弹窗提醒可能成为环境干扰分散被试注意力，而“虚拟人”容易破坏被试作答体验等。

为了避免或减轻控制方法的副作用、反作用，开发更加有效的控制方法，必须回答“控制方法为何有效”的问题。为此，未来研究可以采取一定技术手段(如眼动、脑电等)对被试问卷作答过程进

行深入细致的监控与探索, 丰富、完善不认真作答产生机制及影响因素的相关理论, 并结合这些理论解释产生副作用、反作用的原因, 在此基础上对控制方法进行优化与开发。

另外, 未来研究可以对已有方法进行系统梳理, 分析现有控制方法的具体作用。已有研究常通过实验组和对照组在若干不认真作答反应识别指标上的差异, 对控制方法是否有效做出回应, 但是许多控制方法仅对某些识别指标有作用。因此未来研究可以通过实验设计对各方法的实际效果进行检验与比较, 并结合不认真作答的产生机制解释这些方法降低了何种类型的不认真作答, 为研究者和实践者在选用时提供参考。

### 4.2 探究不认真作答识别指标的跨情境适用性, 开发新方法

已有识别指标多基于人格量表或认知测验开发, 这两类问卷具有题目较多、得分呈正态分布等特点, 因此许多指标在这些问卷情境中有更好的适用性。例如, 问卷越长, 就有越多的题目能用来计算奇偶一致系数、正/反向题目对相关, 得到的系数也更加稳定; 在得分呈正态分布时, 马氏距离、lz 指数等指标也更加有效。

而态度和行为调查这两类社会科学领域同样常见的问卷可能不满足上述特征, 这会造成识别指标有效性下降。例如, 在许多态度问卷中, 正常被试倾向于给出 4 分或 5 分(以五点计分的李克特量表为例), 总体得分呈负偏态, 而一些不认真作答的被试则可能在所有题目上均给 5 分。在这种情况下, 由于个体作答内部差异减小, 许多个体一致性分析指标的效果下降, 同时由于与正常被试的差异较小, 奇异值分析指标有效性也可能下降。

因此, 未来研究需要重点关注不同指标的跨情境适用性。对态度和行为调查而言, 一方面, 结合现有各指标的特点, 组合使用多个指标, 以应对单一指标识别效果不佳的问题。但当指标联合使用时, 要对这些指标各自能识别什么样的不认真作答模式有更清楚的认识, 进而针对各类型的不认真作答模式, 有选择地使用若干相应指标。另一方面, 可以开发新指标, 以应对已有指标不适用的问题, 尤其可以关注个人拟合指数、机器学习的应用, 相较于人总相关系数等传统方法, 这些方法在奇异值识别上更加精准。

另外需要注意, 现有研究多采用模拟研究的方式判断识别指标的有效性, 但现有模拟数据的参数特征可能不适用于态度和行为调查问卷中, 因此未来研究可多利用态度和行为调查的真实数据, 以提高研究的生态效度与研究结果的推广性。

### 4.3 局部不认真作答的识别与处理

尽管已有研究常将被试做“认真作答”与“不认真作答”的区分, 但真实作答情境中, 除了完全不认真的被试外, 也有一部分被试仅在部分题目中作答不认真。例如, 当问卷较长时, 被试更容易在中间或后半部分因疲劳或失去兴趣从而表现出不认真作答(Baer et al., 1997; Berry et al., 1992; Meade & Craig, 2012)。当局部不认真出现时, 嵌入量表的错答次数、个体一致性指标、奇异值分析指标均可能介于完全认真与完全不认真的被试之间, 与完全认真作答的相似性取决于其局部不认真的比例。这种情况下, 通过已有指标可能难以将其识别出来。目前, 对此情况仅 Dunn 等人(2018)指出, 可以灵活地选择部分连续题目, 计算作答标准差, 探测被试在选中题目上是否认真作答。例如, 当问卷较长时, 可以在较为靠后的位置选择若干题目, 判断哪些被试因疲劳等原因出现局部不认真。但是, 被试未必都在这一部分才出现不认真作答。特别是电子问卷兴起后, 被试的作答环境无法控制, 被试可能在任何作答时间内受到外界干扰。因此, 如何采用更灵活的手段识别被试不认真作答的部分, 可成为未来研究的方向之一。

此外, 当成功识别出被试不认真作答的部分时, 对这部分数据的处理也有待进一步研究。若删除该被试的全部数据, 则是对有效数据的浪费; 但仅仅剔除不认真作答的数据, 又会产生数据非随机缺失的风险。即使能够排除非随机缺失情况, 不认真作答的数据也并非缺失数据, 而是不够准确的数据, 它同样代表了被试的部分倾向, 因此是否利用插补处理、以及用何种插补方法都值得进一步探讨。

## 5 小结

被试不会在任何时候都认真思考并给出可靠的答案。研究者与实践者对这一现象更不能盲目乐观或选择性无视, 而应当在利用问卷收集数据时采取有效措施尽可能控制不认真作答的产生, 并在数据清理阶段通过一定技术手段识别并剔除

这类噪音数据，使得数据尽可能真实、准确，以便后续得到可靠的分析结果。

## 参考文献


车文博. (2001). *心理咨询大百科全书*. 杭州：浙江科学技术出版社.

刘蔚华，陈远. (1991). *方法大辞典*. 济南：山东人民出版社.

牟智佳. (2017). MOOCs 学习参与度影响因素的结构关系与效应研究——自我决定理论的视角. *电化教育研究, 38*(10), 37–43.

王俪嘉，朱德全. (2009). 中小学教师对待公开课态度的调查研究. *上海教育科研,* (8), 28–31.

卫旭华，张亮花. (2019). 单题项测量:质疑、回应及建议. *心理科学进展, 27*(7), 1194–1204.

姚成，龚毅，濮光宁，葛文龙. (2012). 学生评教异常数据的筛选与处理. *牡丹江师范学院学报(自然科学版)* (3), 7–8.

郑云翔，杨浩，冯诗晓. (2018). 高校教师信息化教学适应性绩效评价研究. *中国电化教育,* (2), 21–28.

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research, 29*(3), 497–519.

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68*(1), 139–151.

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53*(2), 182–200.

Beach, D. A. (1989). Identifying the random responder. *The Journal of Psychology, 123*(1), 101–103.

Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment, 4*(3), 340.

Börger, T. (2016). Are fast responses more random? Testing the effect of response time on scale in an online choice experiment. *Environmental and Resource Economics, 65*(2), 389–413.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*(2), 218.

Burns, G. N., Christiansen, N. D., Morris, M. B., Periard, D. A., & Coaster, J. A. (2014). Effects of applicant personality on resume evaluations. *Journal of Business and Psychology, 29*(4), 573–591.

Carrier, L. M., Cheever, N. A., Rosen, L. D., Benitez, S., & Chang, J. (2009). Multitasking across generations: Multitasking choices and difficulty ratings in three generations of Americans. *Computers in Human Behavior, 25*(2), 483–489.

Cialdini, R. B. (2001). Harnessing the science of persuasion. *Harvard Business Review, 79*(9), 72–81.

Cibelli, K. L. (2017). *The effects of respondent commitment and feedback on response quality in online surveys.* (Unpublished doctorial dissertation), University of Michigan, Ann Arbor.

Costa Jr, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE Handbook of Personality Theory and Assessment: Personality Measurement and Testing* (pp. 179–198). London: SAGE Publications Ltd.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*(4), 596–612.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology, 67*(2), 309–338.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*(2), 171–181.

Drasgow, F., Levine, M. V. & Williams, E. A. (1985) Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology 38*, 67–86.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology, 33*(1), 105–121.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224–247.

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research, 15*(2), 195–219.

Fang, J. M., Prybutok, V., & Wen, C. (2016). Shirking behavior and socially desirable responding in online surveys: A cross-cultural study comparing Chinese and American samples. *Computers in Human Behavior, 54*, 310–317.

Fang, J. M., Wen, C., & Prybutok, V. (2014). An assessment of equivalence between paper and social media surveys: The role of social desirability and satisficing. *Computers in Human Behavior, 30*, 335–343.

Francavilla, N. M., Meade, A. W., & Young, A. L. (2019). Social interaction and internet-based surveys: Examining the effects of virtual and in-person proctors on careless response. *Applied Psychology, 68*(2), 223–249.

García, A. A. (2011). Cognitive interviews to test and refine questionnaires. *Public Health Nursing, 28*(5), 444–450.

Gough, H. G., & Bradley, P. (1996). *The California psychological inventory™ manual: Third edition.* Palo Alto, CA: Consulting Psychologists Press.

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural differences in careless responding. Journal of *Cross-Cultural Psychology, 50*(3), 336–357.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60−90). Princeton, NJ: Princeton University Press.

He, J., & van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences, 55*(7), 794–800.

He, J., & van de Vijver, F. J. R. (2015a). Effects of a general response style on cross-cultural comparisons: Evidence from the teaching and learning international survey. *Public Opinion Quarterly, 79*(S1), 267–290.

He, J., & van de Vijver, F. J. R. (2015b). Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences, 81*, 129–134.

He, J., & van de Vijver, F. J. R. (2016). Response styles in factual items: Personal, contextual and cultural correlates. *International Journal of Psychology, 51*(6), 445–452.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (2000). *Understanding robust and exploratory data analysis*. New York, NY: John Wiley.

Höhne, J. K., & Schlosser, S. (2018). Investigating the adequacy of response time outlier definitions in computer-based web surveys using paradata SurveyFocus. *Social Science Computer Review, 36*(3), 369–378.

Holtzman, N. S., & Donnellan, M. B. (2017). A simulator of the degree to which random responding leads to biases in the correlations between two individual differences. *Personality and Individual Differences, 114*, 187–192.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114.

Huang, J. L., Liu, M. Q., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*(3), 828–845.

Jackson, D. N. (1976). *The appraisal of personal reliability*. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

Jackson, D. N. (1977). *Jackson vocational interest survey: manual.* Port Huron, MI: Research Psychologists Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1), 103–129.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality. *Organizational Research Methods, 18*(3), 512–541.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298.

Kountur, R. (2016). Detecting careless responses to self-reported questionnaires. *Eurasian Journal of Educational Research,* (64), 307–318.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003–1020.

Levine, M. V., & Rubin, D. B. (1979) Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics 4*, 269–290.

Lloyd, K., & Devine, P. (2010). Using the internet to give children a voice: An online survey of 10- and 11-year-old children in Northern Ireland. *Field Methods, 22*(3), 270–289.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*, 49–55.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83.

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious

and random responders. *Personality and Individual Differences, 84*, 79–83.

Mayerl, J. (2013). Response latency measurement in surveys: Detecting strong attitudes and response effects. *Survey Methods: Insights from the Field*. Retrieved from https://surveyinsights.org/?p=1063

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*(3), 450–470.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135.

Melipillán, E. R. (2019). *Careless survey respondents: Approaches to identify and reduce their negative impact on survey estimates.* (Unpublished doctorial dissertation), University of Michigan, Ann Arbor.

Nguyen, H. L. T. (2017). *Tired of survey fatigue? Insufficient effort responding due to survey fatigue* (Unpublished master's thesis), Middle Tennessee State University, Murfreesboro.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality, 63*, 1–11.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872.

Revilla, M., & Ochoa, C. (2015). What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science Computer Review, 33*(1), 97–114.

Rousseau, B., & Ennis, J. M. (2013). Importance of correct instructions in the tetrad test. *Journal of Sensory Studies, 28*(4), 264–269.

Schneider, S., May, M., & Stone, A. A. (2018). Careless responding in internet-based quality of life assessments. *Quality of Life Research, 27*(4), 1077–1088.

Schnell, R. (1994). *Graphisch gestützte datenanalyse [Graphically supported data analysis]*. München, Germany: Oldenbourg.

Soland, J., Wise, S. L., & Gao, L. Y. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education, 32*(2), 151–165.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties [Comparability of individual test performance]*. Lisse, Netherlands: Swets & Zeitlinger.

Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician, 35*(4), 234–242.

Wang, C., & Xu, G. J. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology, 67*(2), 231–263.

Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior, 48*, 554–568.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61.

Wise, S. L., & Demars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38.

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 186–191.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology, 22*(1), 51–68.

Zhang, C. (2013). *Satisficing in web surveys: Implications for data quality and strategies for reduction.* (Unpublished doctorial dissertation). University of Michigan, Ann Arbor.

Zhang, C., & Conrad, F. G. (2018). Intervening to reduce satisficing behaviors in web surveys. *Social Science Computer Review, 36*(1), 57–81.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics, 36*(2), 186–212.

# Preventing and detecting insufficient effort survey responding

ZHONG Xiaoyu, LI Mingyao, LI Lingyan

(*Collaboration Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing* 100875, *China*)

**Abstract:** Surveys are commonly used in psychological and educational research. Insufficient effort response (IER), as one source of invalid response data, is somewhat prevalent due to the low-stakes nature of the majority of surveys, which often leads to statistically significantly biased estimates and invalid inferences. The current literature shows: (a) IER is commonly believed to be caused by some inner causes, (e.g., low motivation), showing as specific patterns, (e.g., random responding); (b) The most common methods to prevent IER include reducing task difficulty and increasing respondents' motivation; (c) Current detection methods fall into three main categories, which are proactive approaches/ direct screening methods, response patterns analysis, and response time analysis. Recommendations for future research directions and practitioners are (a) deepening the investigation on IER mechanism and improving the preventing methods, (b) examining the effectiveness of IER identification methods' applicability of cross-situation and developing new approaches, and (c) delving into the identification and treatment of partial IER.

**Key words:** insufficient effort responding (IER), data screening, invalid response, survey and questionnaire design & construction